Assignment - 2
Machine Learning

Akshay Tambe (apt321@nyu.edu)
Snahil Singh (ss11381@nyu.edu)

Q1.

binary (2-class) = +/-
Predict label of test example x
correct label → r
predicted label → y

TP → $r = +, y = +$ TN → $r = -, y = -$
FP → $r = -, y = +$ FN → $r = +, y = -$

Msg. No	r	y
1	−	+
2	+	−
3	+	−
4	−	−
5	−	−
6	+	−

1 FP
3 FN
0 TP
2 TN

Confusion matrix
Predicted (y)

Correct (r)

	+	−
+	TP	FN
−	FP	TN

	+	−
+	0	3
−	1	2

a) $TPR = \dfrac{\#\text{ TP examples}}{\#\text{ ex. where correct label } (+)} = \dfrac{\#TP}{\#TP + \#FN} = 0$

b) $FPR = \dfrac{\#\text{ FP examples}}{\#\text{ examples whose correct label } (-)} = \dfrac{\#FP}{\#FP + \#TN} = \dfrac{1}{1+2} = 1/3$

c. Accuracy $= \frac{TN+TP}{TN+TP+FN+FP} = \frac{41+56}{56+2+41+1} = \frac{97}{100}$

		Predicted +	Predicted −
correct	+	56	2
	−	1	41

Q2. Binary classification (2-classes → C_1 & C_2)

Attributes → x_1, x_2

LDF → $g_1(x_1, x_2) = 5x_2 + 3x_1 - 4$ for C_1

$g_2(x_1, x_2) = -3x_2 + 2x_1 - 6$ for C_2

New sample → $\underbrace{g_1(x_1, x_2) \cdot g_2(x_1, x_2)}_{\text{Bigger value} \to C_1}$

Tie → assign C_2

Single Discriminant function → $g(x_1, x_2)$

$= w_2 x_2 + w_1 x_1 + w_0 x_0$

where $C_1 \to g(x_1, x_2) > 0$

$C_2 \to$ otherwise

$g(x_1, x_2)$ is linear, the decision surface is a hyperplane. Therefore, it divides the feature space into decision region. R_1 for C_1 & R_2 for C_2 separated by boundaries. When there are 2-classes, we can define a single discriminant function as - $g(x_1, x_2) = g_1(x_1, x_2) - g_2(x_1, x_2)$

$= (5+3)x_2 + (3-2)x_1 + (-4+6)$

$= 8x_2 + x_1 + 2$

where $C_1 \to g(x_1, x_2) > 0$

$C_2 \to$ otherwise

Q3. spam (+) Not spam (−) ← 2 classes

FP: NS. mail (−) ⟶ S(+)

FN: S mail (+) ⟶ NS(−)

a) $\hat{P}(-) = a * P(C = +|x)$

$\hat{P}(+) = b * P(C = -|x)$

$$P(\text{spam}|x_1, x_2) = \frac{1}{1+e^{-(3x_2 - 2x_1 + 1)}}$$

$x_1 = 3,\ x_2 = 2$

FN cost : 5 , FP cost : 2

$$P(+|x_1, x_2) = \frac{1}{1+e^{-(-6-6+1)}} = 1/_{1+e^{-1}}$$

$$P(-|x_1, x_2) = 1 - 1/_{1+e^{-1}}$$

Cost with $\hat{P}(NS) = a * P(c = +|x) = 5\left(1/_{1+e^{-1}}\right) = \frac{5e}{1+e}$

Cost with $\hat{P}(S) = b * P(c = -|x) = 2\left(1 - 1/_{1+e^{-1}}\right) = \frac{2}{e+1}$

Cost with $\hat{P}(NS)$ > Cost with $\hat{P}(S)$, so predicting spam has lower risk.

Q4. D → Gaussian Distb.

$\downarrow$,

$X\,]_{t=1}^{N}$

$$E(\text{mean}) = \frac{\Sigma_{x \in X}\, x^t}{N}$$

but we took, $\frac{\Sigma_{x \in X}\, x^t}{N+1}$

Bias of estimator $= E(\text{mean}) - \text{actual}$

$$= \frac{\Sigma_{x \in X}\, x^t}{N+1} - \frac{\Sigma_{x \in X}\, x^t}{N} = \frac{-\Sigma_{x \in X}\, x^t}{N(N+1)}$$

Scanned by CamScanner

b) If D is an exponential distribution, bias won't change because,

Let $\{x_i\}$ be iid random var exp. distributed as $x_i \sim Exp(\lambda)$

likelihood function with $X = \{x_i\}$ can be,

$$p(X) = \prod_{i=1}^{N} \lambda e^{-\lambda x_i}.$$

with likelihood as,

$$L(\lambda, X) = N \log \lambda - \lambda \sum_{i=1}^{N} x_i$$

MLE estimate with $\hat{\lambda} = \underset{\lambda}{\arg\min}\ L(\lambda, X) \Rightarrow \hat{\lambda} = 1/\bar{x}$

where, $\bar{x}$ is mean of $\{x_i\}$, $\bar{x} = 1/N \sum_{i=1}^{N} x_i$

Thus, $\hat{\mu} = \bar{x}$ & we took $N = N+1$, we'll get same bias

Q5. $\hat{\mu} = \dfrac{\sum_{t=1}^{N} x^t}{N}$ $\hat{\Sigma} = \dfrac{\sum (x^t - \hat{\mu})(x^t - \hat{\mu})^T}{N}$

For + class,

x_1	x_2	Label
2.7	4.8	+
3.2	5.1	+
−0.4	−0.3	+

$\mu_{x1} = \dfrac{2.7 + 3.2 - 0.4}{3} = 1.833$ $\mu_{x_2} = \dfrac{4.8 + 5.1 - 0.3}{3} = 3.2$

For − class,

x_1	x_2	Label
~~0.4~~	~~0.3~~	
0.6	0.5	−
1.8	2.8	−
2.1	4.3	−

$\mu_{x1} = \dfrac{0.6 + 1.8 + 2.1}{3} = 1.5$ $\mu_{x_2} = \dfrac{0.5 + 2.8 + 4.3}{3} = 2.533$

For + class $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^{3}(x_{1i}-\mu_{x1})^2}{N} & \frac{\sum(x_{1i}-\mu_{x1})(x_{2i}-\mu_{x2})}{N} \\ \frac{\sum(x_{1i}-\mu_{x1})(x_{2i}-\mu_{x2})}{N} & \frac{\sum_{i=1}^{3}(x_{2i}-\mu_{x2})^2}{N} \end{bmatrix}$

$= (2.7-1.8333)^2 + (3.2-1.8333)^2 + (-0.4-1.8333)^2/3 = \sigma_1^2$

$= 2.53\cancel{3}5$

$= (4.8-3.2)^2 + (5.1-3.2)^2 + (-0.3-3.2)^2/3 = \sigma_2^2 = 6.14$

$\sigma_{12} = \sum_{i=1}^{3}(x_{1i}-\mu_{x1})(x_{2i}-\mu_{x2})/N$

$\sigma_{12} = \sigma_{21} = [(2.7-1.8333)(4.8-3.2) + (3.2-1.8333)(5.1-3.2) +$
$(-0.4-1.8333)(-0.3-3.2)]/3$

$= 1.3872 + 2.5973 + 7.8155 = \frac{11.8}{3} = 3.933$

For − class $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^{3}(x_{1i}-\mu_{x1})^2}{N} & \frac{\sum_{i=1}^{3}(x_{1i}-\mu_{x1})(x_{2i}-\mu_{x2})}{N} \\ \frac{\sum_{i=1}^{3}(x_{1i}-\mu_{x1})(x_{2i}-\mu_{x2})}{N} & \frac{\sum_{i=1}^{3}(x_{2i}-\mu_{x2})^2}{N} \end{bmatrix}$

$\sigma_1^2 = [(0.6-1.5)^2 + (1.8-1.5)^2 + (2.1-1.5)^2]/3 = 0.42$

$\sigma_2^2 = [(0.5-2.53)^2 + (2.8-2.53)^2 + (4.3-2.53)^2]/3 = 2.4422$

$\sigma_{12} = \sigma_{21} = [(0.6-1.5)(0.5-2.53) + (1.8-1.5)(2.8-2.53) +$
$(2.1-1.5)(4.3-2.53)]/3 = 0.99$

as $\mu_{x1} = 1.5$, $\mu_{x2} = 2.53$

For +ve class

$\Sigma = \begin{bmatrix} 2.53\cancel{3}5 & 3.933 \\ 3.933 & 6.14 \end{bmatrix}$

$\mu_{x1} = 1.833$
$\mu_{x2} = 3.2$

For −ve class

$\Sigma = \begin{bmatrix} 0.42 & 0.99 \\ 0.99 & 2.4422 \end{bmatrix}$

$\mu_{x1} = 1.5$
$\mu_{x2} = 2.533$

b) $pdf = \frac{1}{(2\pi)^{d/2} \cdot |\Sigma|^{1/2}} \cdot e^{-1/2 (x-\mu)^T \Sigma^{-1} (x-\mu)}$

$x = \begin{pmatrix} 1.6 \\ 2.3 \end{pmatrix}$ $\log p(x|+)$ and $\log p(x|-)$

$d = 2$, $|\Sigma_+| = 0.096411$

~~log~~ . ϕ $\Sigma_+^{-1} = \begin{bmatrix} 63.68567 & -40.7941 \\ -40.7941 & 26.29368 \end{bmatrix}$ $x = \begin{bmatrix} 1.6 \\ 2.3 \end{bmatrix}$ $\mu = \begin{bmatrix} 1.833 \\ 3.2 \end{bmatrix}$

$\log p(x|+) = \log \left[e^{-1/2} \cdot \left[(1.6-1.833)\ (2.3-3.2) \right] \begin{bmatrix} 63.68567 & -40.7941 \\ -40.7941 & 26.29368 \end{bmatrix} \begin{bmatrix} 1.6-1.833 \\ 2.3-3.2 \end{bmatrix} \right]$

$- \log \left((2(3.14))^{2/2} \cdot (0.096411)^{1/2} \right)$

$= \log_e e^{[-1/2 (7.6468)]} - 1.95\cancel{8}093 = \cancel{-0.41719}$

$= \cancel{-11.8811}$ $\quad -\frac{1}{2}(7.64626) - 1.95093 = -5.77406$

$|\Sigma_-| = 0.045624$

$\Sigma_-^{-1} = \begin{bmatrix} 53.5288 & -21.6991 \\ -21.6991 & 9.2056 \end{bmatrix}$ $x = \begin{bmatrix} 1.6 \\ 2.3 \end{bmatrix}$ $\cancel{\mu = \begin{bmatrix} 1.833 \\ 3.2 \end{bmatrix}}$ $\mu = \begin{bmatrix} 1.5 \\ 2.533 \end{bmatrix}$

$\log p(x|-) = \log \left[e^{-1/2} \cdot \left[(1.6-1.5)\ (2.3-2.533) \right] \begin{bmatrix} 53.5288 & -21.699 \\ -21.699 & 9.205 \end{bmatrix} \begin{bmatrix} 1.6-1.5 \\ 2.3-2.533 \end{bmatrix} \right]$

$- \log \left[(2.3.14)^{2/2} \times (0.045624)^{1/2} \right] = \log(e^{-1/2} \cdot (2.046191)) - \log(1.34207)$

$= \cancel{-1.1948} \cancel{-0.07823} = -\frac{1}{2}(2.046191) - \log(1.34207)$

$= -1.02309 - 0.2942 = -1.3$

Using ML hypothesis, $\underset{\theta}{\operatorname{argmax}} [\log P(x|+), \log P(x|-)]$

we find that $\log P(x|-) > \log p(x|+)$

$\therefore$ it belongs to −ve class

Scanned by CamScanner

c) We can have one covariance matrix in cases where attributes do not strictly adhere to a class (positive or negative classes here), so having one matrix won't impact much in classifying test data.

For eg:- In cases of strings, punctuation marks → '.', ',', ';' are there, we can think to have one covariance matrix for such attributes.

However, cases where attributes strictly more or less adhere to a class & plays an important role in classifying data & there's a difference in the sets of attributes that define a positive class with ones that define the negative class, we need to compute different covariance matrices, because each class'll ~~have~~ tend towards a different matrix.

86. Using simple linear regression

$$A = \begin{bmatrix} N & \sum_t x^t \\ \sum_t x^t & \sum_t (x^t)^2 \end{bmatrix} \quad W = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad Y = \begin{bmatrix} \sum_t r^t \\ \sum r^t x^t \end{bmatrix}$$

Ar –

X	Y	XY	X^2
1050	57	59850	1102500
428	28	11984	183184
362	26	9412	131044
529	40	21160	279841
790	60	47400	624100
401	22	8822	160801
380	38	14440	144400
1454	110	159940	2114116
1127	100	112700	1270129
700	46	32200	490000

$$A = \begin{bmatrix} 10 & 7221 \\ 7221 & 6500115 \end{bmatrix} \quad Y = \begin{bmatrix} 527 \\ 477908 \end{bmatrix}$$

$$W = A^{-1}Y \Rightarrow A^{-1} = \begin{bmatrix} 0.5055186 & -0.00056158 \\ -0.00056158 & 7.777072390 \end{bmatrix}$$

$$w = A^{-1}Y = \begin{bmatrix} 0.505518 & -0.00056158 \\ -0.00056158 & 7.777072390 \times 10^{-7} \end{bmatrix} \begin{bmatrix} 527 \\ 477908 \end{bmatrix} = \begin{bmatrix} -1.9763923 \\ 0.0757185 \end{bmatrix}$$

$$\text{line } g(x|w_1, w_0) = w_1 x + w_0 = \cancel{(-1.9763923)\, x + 0.0757185}$$

$$= (0.0757185)\,x - 1.9763923$$

b) Predicted no of stories i.e 475 ft tall

$= (0.0757185)475 - 1.9763923$

$= 35.9662875 - 1.9763923 = 33.9899952$

$\simeq 34$ Stories

Q.7. K-NN → Imp to scale attributes

$$\text{Scale}(x_i) = \frac{x_i - x_i\ min}{x_i\ max - x_i\ min}$$

$x_i\ min$ → min value of x_i in training set

$x_i\ max$ → max value of x_i in training set

a) Original dataset

x_1	x_2	label
2.5	~~42~~42	+
3.8	51	+
-0.3	-1	+
0.7	3	−
1.6	26	−
2.3	41	−

Scaled Dataset

x_1	x_2	label
0.68	0.83	+
1.00	1.00	+
0.00	0.00	+
0.24	0.08	−
0.46	0.52	−
0.63	0.81	−

$x_1^{min} = -0.3$, $x_1^{max} = 3.8$ $\quad x_1^{max} - x_1^{min} = 4.1$

$x_2^{min} = -1$, $x_2^{max} = 51$ $\quad x_1^{max} - x_1^{min} = 52$

b) x_i^{min} → min value of x_i in training set

x_i^{max} → max value of x_i in training set

$x_1^{min} = -0.3$ $\quad$ $x_1^{max} = 3.8$.

$x_2^{min} = -1$ $\quad$ $x_2^{max} = 51$

Test data

$x = \begin{bmatrix} 3.9 \\ 4 \end{bmatrix}$

Scaled test data

$x = \begin{bmatrix} 1.02 \\ 0.096 \end{bmatrix}$

	x_1	x_2	(1.02, 0.096)
+	0.68	0.83	$\sqrt{(0.68-1.02)^2 + (0.096-0.83)^2} = 0.808$
+	1.00	1.00	0.904
+	0.00	0.00	1.626
−	0.24	0.08	0.78
−	0.46	0.52	0.702 . * Min Ed K=1
−	0.63	0.81	0.813 .

Predicted label K = 1 = (−)